

Random threshold for linear model selection, revisited.

Merlin Keller¹, Marc Lavielle²,

¹ EDF R&D, Chatou, France

² INRIA Saclay, France and University of Paris Sud, Orsay, France

Contents

1	Introduction	2
2	Original random thresholding procedure	3
2.1	Testing the presence of significant coefficients	3
2.2	Selecting the significant coefficients	6
3	Extensions and asymptotic properties	7
3.1	Unknown distribution extension	7
3.2	Varying window extension	8
3.3	Asymptotic properties	10
4	Simulation study	11
4.1	Results with known null distribution	11
4.2	Results with unknown null distribution	12
5	Application to fMRI data analysis	14
5.1	Data acquisition and preprocessings	15
5.2	Individual subject activation map	15
5.3	Group activation map	16
6	Conclusion	16
A	Proof of the Theorem 2 for the varying window extension	19
B	Details of the EM algorithm for the two-class GMM with a zero-mean class	20

Abstract In [9], a random thresholding method is introduced to select the significant, or non null, mean terms among a collection of independent random variables, and applied to the problem of recovering the significant coefficients in non ordered model selection. We introduce a simple modification which removes the dependency of the proposed estimator on a window parameter while maintaining its asymptotic properties. A simulation study suggests that both procedures compare favorably to standard thresholding approaches, such as multiple testing or model-based clustering, in terms of the binary classification risk. An application of the method to the problem of activation detection on functional magnetic resonance imaging (fMRI) data is discussed.

1 Introduction

In [9], the following model is considered:

$$Y_i = \mu_i + \varepsilon_i, \quad i = 1, \dots, n, \quad (1)$$

where μ_i are unknown constants, some of which are zero, and ε_i are independent, identically distributed (iid) zero-mean random variables, with known cumulative distribution function (cdf) F_ε .

Within this model, the problem of selecting the significant coefficients $\mu_i \neq 0$, based on the observations Y_i is studied. Such a problem arises in many different application areas, such as genomics [7], or neuroimaging [6], to cite just a few.

Many methods have been proposed to perform this task. Multiple testing procedures for instance (see [7] for an overview of existing methods), have been developed to control a certain type I error rate, such as the familywise error rate (FWER), or the false discovery rate (FDR) [2] at a user-fixed level. It can be argued however that the choice of a level, which ultimately defines the subset of selected coefficients, is arbitrary, as there is no safe guideline to what an ‘optimal’ level of false detections should be.

An alternative, that allows to control both type I and type II error rates, consists in fitting a mixture model to the data, with one class for the null (zero-mean) data, and one, or more, for the non-null data. A detection threshold can then be derived, which minimizes a certain classification risk, such as the binary risk, associated to the 0-1 loss function, resulting in a ‘naive Bayes’ classifier [10]. The main difficulty of this approach lies in the choice of a distribution for the non-null data, which may influence significantly the resulting classifier. Many authors have proposed to deal with this issue through model selection techniques (see [8, 11, 5] for instance), however it remains an open-ended problem.

In view of these difficulties, the random threshold (RT) approach introduced in [9] appears as a promising candidate, since it does not require the specification of a type-I error level, nor of a model for the non-zero mean observations. The principle of RT lies in estimating the number of significant coefficients, based on a random centering of the partial sums of the ordered observations. Because it relies on as little assumptions as possible, we expect RT to be more robust than the above-mentioned approaches.

However, to date very little is known concerning the classification performances of RT procedures; [9] essentially gives a minimal separation speed between null and non-null data for the method to attain perfect classification asymptotically. Furthermore, the algorithm described therein still depends on a window parameter, which may have some influence in presence of noisy data.

This article describes a simple modification of the RT procedure, which removes its dependency on the window parameter, while maintaining its asymptotic properties. We then study the classification performances of both techniques using numerical experiments, in comparison to the above-mentioned standard approaches.

The rest of this article is organized as follows. In Section 2, the original RT method is reviewed. The variable window extension is introduced in Section 3. In Section 4 the results of numerical experiments are presented, which show the good properties of RT in terms of classification. An application to fMRI data analysis is discussed in Section 5, and we conclude in Section 6 by considering the perspectives open by this work.

2 Original random thresholding procedure

2.1 Testing the presence of significant coefficients

We start by recalling how the presence of non-zero means is tested, that is, how the null hypothesis $\mathcal{H}_0 : \mu_i \equiv 0$ is tested in [9]. This is done by comparing the cumulative sums of the ordered observations to their conditional expectations under \mathcal{H}_0 , according to the following steps:

1. Order the observations $|Y_{(1)}| \geq |Y_{(2)}| \geq \dots \geq |Y_{(n)}|$
2. For $i = 1, \dots, n$, let $X_{(i)} = -\log(1 - F_{|\epsilon|}(|Y_{(i)}|))$
3. Let $T_j = \sum_{i=1}^j X_{(i)}$ and $Q_j = \mathbb{E}_{\mathcal{H}_0}(T_j | T_n)$
4. Define the test statistic $D_n = \max_j |T_j - Q_j| / \sqrt{n}$. The null hypothesis is rejected if $D_n > d_\alpha$, with d_α such that $\mathbb{P}_{\mathcal{H}_0}(D_n > d_\alpha) \leq \alpha$.

Note that the cumulative sums are not computed directly from the ordered observations, but from the transforms $X_{(1)}, \dots, X_{(n)}$ which, under \mathcal{H}_0 , are an ordered series of $\mathcal{E}(1)$ random variables. The conditional expectations $Q_j = \mathbb{E}_{\mathcal{H}_0}(T_j | T_n)$ can then be computed using the following result (see [9] for details);

Theorem 1 (Conditional expectations of ordered exponential variables) *Under \mathcal{H}_0 , the $X_{(i)}$ are an ordered series of $\mathcal{E}(1)$ random variables. It follows that:*

- i) $\mathbb{E}_{\mathcal{H}_0}(X_{(i)}) = \sum_{\ell=i}^n \frac{1}{\ell}$
- ii) $\mathbb{E}_{\mathcal{H}_0}(T_j) = j + j \sum_{\ell=j}^n \frac{1}{\ell}$
- iii) $\mathbb{E}_{\mathcal{H}_0}(T_j | T_n) = \frac{\mathbb{E}_{\mathcal{H}_0}(T_j)}{\mathbb{E}_{\mathcal{H}_0}(T_n)} T_n.$

Furthermore, for $n \geq 100$, it is shown in [9] using Monte-Carlo simulations that:

$$\mathbb{P}_{\mathcal{H}_0}(D_n > 0.65) \approx 0.05, \quad (2)$$

which provides an approximate calibration for the above test. It is illustrated in Figure 1, on a dataset of $n = 500$ observations Y_i simulated according to model (1), with noise terms ε_i sampled from the $\mathcal{N}(0, 1)$ distribution.

Under the global null \mathcal{H}_0 , that is, when $\mu_i \equiv 0$, (Figure 1, top), the cumulative sums (T_j) are seen to follow closely their conditional expectations (Q_j). Consequently, the resulting test statistic value $D_n = 0.55$, does not exceed the critical value 0.65 given by (2).

In contrast, when we add $n_1 = 100$ non-zero means, all taken equal to $\mu_i \equiv 5$, to the null data (Figure 1, bottom), the T_j 's become substantially larger than their expected values Q_j under \mathcal{H}_0 , resulting in a gap between the corresponding curves. Note that this gap is most significant around $j = 100$, because, for all j , T_j is the sum of the j largest observations (after transformation), containing mostly non-zero means for $j \leq 100$. Consequently, the ensuing test statistic $D_n = 23.44$ is far above the critical value 0.65.

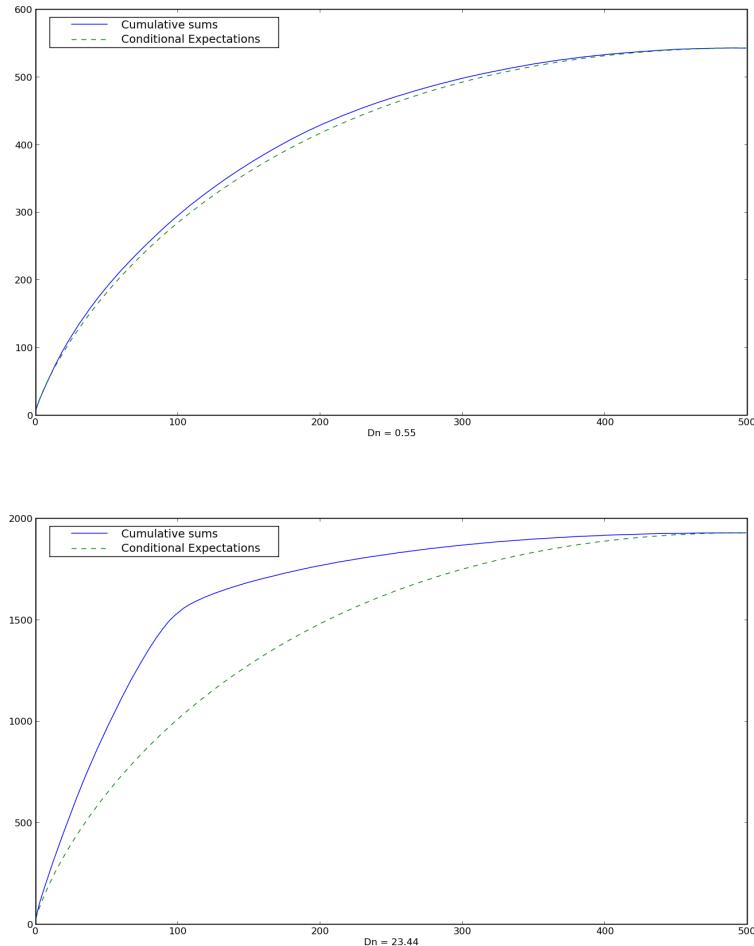


Figure 1: Test of the presence of significant coefficients: the cumulative sums (T_j) and their conditional null expectations (Q_j), under the global null \mathcal{H}_0 (top) and in presence of significant coefficients (bottom).

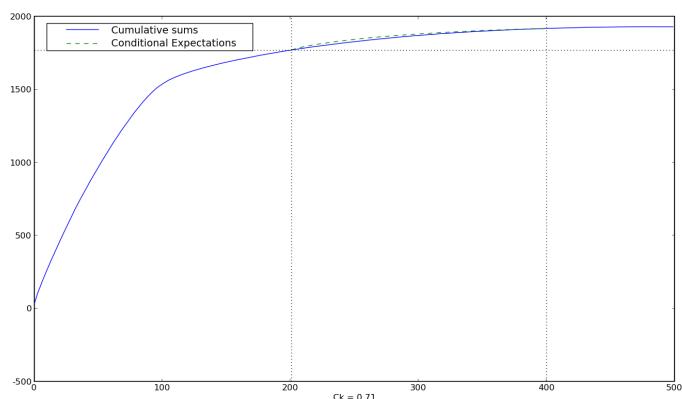
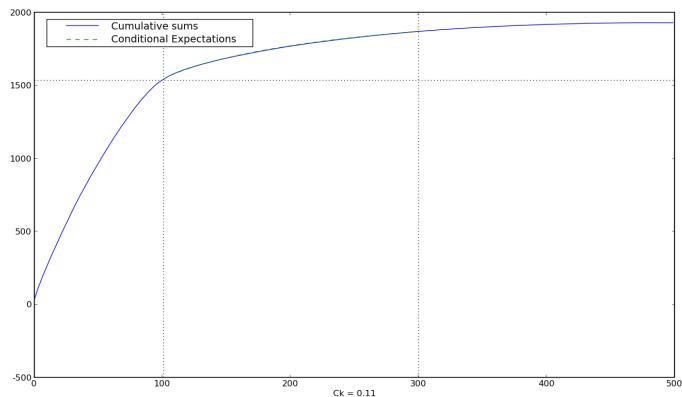
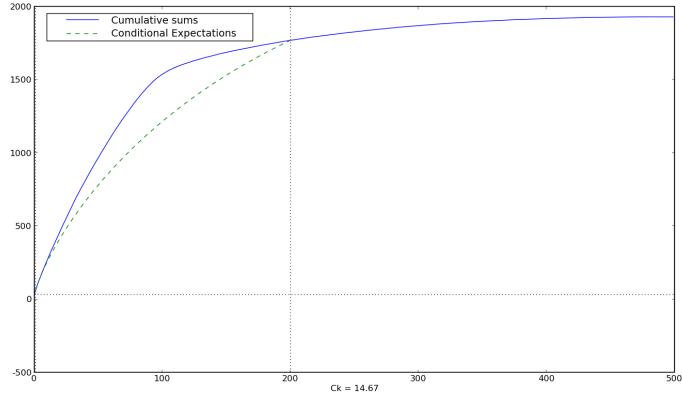


Figure 2: Random threshold procedure: partial cumulative sums $(T_{k,j})_j$ and their conditional null expectations $(Q_{k,j})_j$ for $k = 0$ (top), $k = 100$ (middle), and $k = 200$ (bottom), with window width $K_n = 200$.

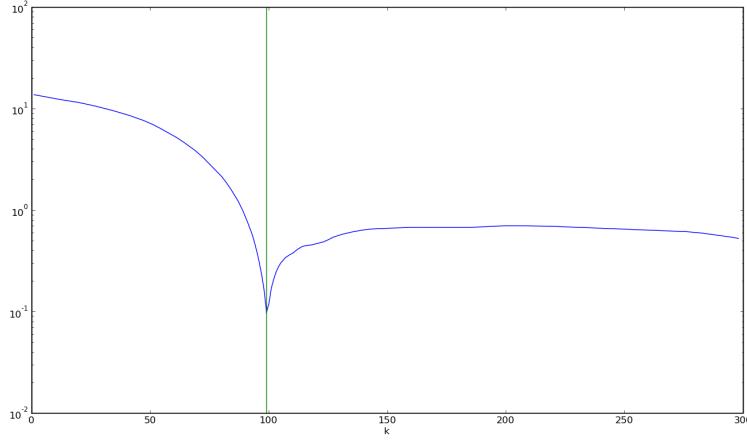


Figure 3: Random threshold procedure: the sequence η_k on a logarithmic scale; its minimum is attained for $\hat{k}_n = 99$.

2.2 Selecting the significant coefficients

Upon rejection of the null hypothesis, the following task consists in selecting the significant coefficients. The procedure for doing so can be interpreted in a data-dependent ‘multiple hypothesis testing’ setting, as described hereafter. Consider the null hypothesis \mathcal{H}_0 , as defined in Section 2.1, and the set of alternative hypotheses:

$$\mathcal{H}_{1,k} : \text{for any } i \leq k, \mu_{(i)} > 0, \text{ and } \mu_{(k+1)} = \dots = \mu_{(n)} = 0.$$

In other terms, $\mathcal{H}_{1,k}$ corresponds to the hypothesis that the k largest observations only have non-zero means. Even though in real-life datasets null and non-null data are never perfectly separated, in general one cannot expect more than to discriminate between such hypotheses in non-ordered model selection. Note that this is equivalent to choosing a certain detection threshold to separate null from non-null data.

Denote \mathbb{E}_k the expectation under $\mathcal{H}_{1,k}$ (instead of $\mathbb{E}_{\mathcal{H}_{1,k}}$). The RT procedure first computes the $X_{(i)}$ ’s using the same steps 1. and 2. as in Section 2.1, then adds the following steps:

3. Let K_n be some positive integer. For $1 \leq k \leq n - K_n$ and $1 \leq j \leq K_n$, compute:

$$\begin{aligned} T_{k,j} &= \sum_{i=k+1}^{k+j} X_{(i)} \\ Q_{k,j} &= \mathbb{E}_k(T_{k,j} | T_{k,K_n}) \\ \eta_k &= \max_{1 \leq j \leq K_n} |T_{k,j} - Q_{k,j}| / \sqrt{n}. \end{aligned}$$

4. Let $\hat{k}_n = \operatorname{argmin}_{1 \leq k \leq n - K_n} \eta_k$.

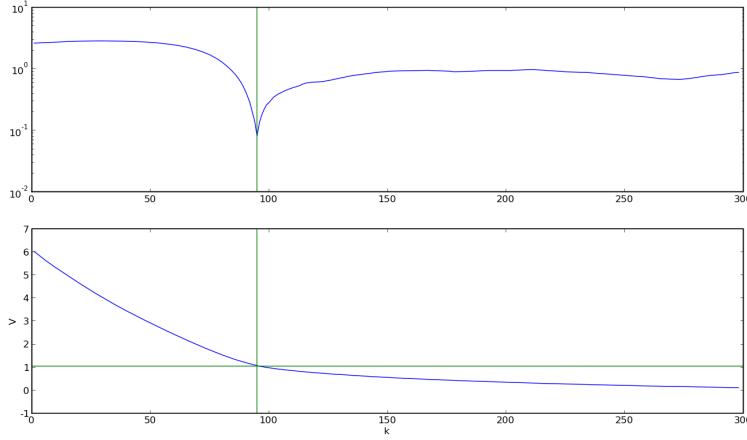


Figure 4: Random threshold method for a Gaussian null distribution with unknown variance σ^2 : the sequence $\eta_k(\hat{\sigma}_k)$ (top) and the sequence $\hat{\sigma}_k$ (bottom).

As in the preceding section, for each k $X_{(k+1)}, \dots, X_{(k+n-k)}$ is an ordered series of $\mathcal{E}(1)$ variables under $\mathcal{H}_{1,k}$, so $Q_{k,j}$ can easily be computed using Theorem 1 (see [9] for further details). Heuristically, the partial cumulative sums $(T_{k,j})$ are compared to their conditional expected values $Q_{k,j}$ under the hypotheses $\mathcal{H}_{1,k}$, for $k = 1, \dots, n - K_n$. The number of significant coefficients is estimated as the index \hat{k}_n corresponding to the minimal gap between $T_{k,j}$ and $Q_{k,j}$, as evaluated by η_k .

This is illustrated in Figure 2, using a dataset simulated exactly as in Section 2.1, that is, with $n_1 = 100$ significant coefficients. Informally, the procedure uses a sliding window with width K_n , and compares the cumulative sums (T_j) within this window to their conditional expectations under the hypothesis that the window contains the K_n largest null terms. For $k = 1$, the window contains in fact mainly significant terms, so that the $T'_{1,j}$ s are well above their expected values, yielding a normalized gap of $\eta_1 = 14.67$. For $k = 100$, the window indeed contains mostly the K_n largest null terms, so $T_{100,j}$ and $Q_{100,j}$ are of the same order, yielding a much smaller gap ($\eta_{100} = 0.11$). Finally, for $k = 200$, the window contains null terms, but not the K_n largest, so the cumulative sums $(T_{k,j})$ become lower than their expected values. Consequently, the gap increases ($\eta_{200} = 0.71$). Figure 3 shows the complete sequence of η_k values, with a clear minimum at $\hat{k}_n = 99$, close to the true number of significant coefficients.

3 Extensions and asymptotic properties

3.1 Unknown distribution extension

The RT method recalled in the previous Sections may be difficult to apply to real-life problems, where the noise distribution F_ϵ is in general unknown. In [9], an extension is proposed to the case where F_ϵ is a parametric distribution $F_\epsilon(\cdot; \theta)$, with θ unknown.

Quite naturally, this consists in estimating θ under each hypothesis $\mathcal{H}_{1,k}$ from the null data $Y_{(k+1)}, Y_{(k+2)}, \dots, Y_{(n)}$, then using this estimate to derive the transforms $X_{(i)}$, for $i = k+1, \dots, k+K_n$. More precisely, having chosen some positive integer K_n , the extension consists in performing for $1 \leq k \leq n - K_n$ the following steps:

1. let $\hat{\theta}_k = \hat{\theta}(Y_{k+1}, \dots, Y_n)$ be an estimate of θ
2. for $i = 1, \dots, n$, let $X_{(i)}(\hat{\theta}_k) = -\log \left(1 - F_{|\epsilon|}(|Y_{(i)}|; \hat{\theta}_k) \right)$,
3. for $1 \leq j \leq K_n$, compute:
 - $T_{k,j}(\hat{\theta}_k) = \sum_{i=k+1}^{k+j} X_{(i)}(\hat{\theta}_k)$
 - $Q_{k,j}(\hat{\theta}_k) = \mathbb{E}_k(T_{k,j}(\hat{\theta}_k) | T_{k,K_n}(\hat{\theta}_k))$
4. Compute $\eta_k(\hat{\theta}_k) = \max_{1 \leq j \leq K_n} |T_{k,j}(\hat{\theta}_k) - Q_{k,j}(\hat{\theta}_k)| / \sqrt{n}$.

Finally, the estimated number of components is given as before by

$$\hat{k}_n = \operatorname{argmin}_{1 \leq k \leq n - K_n} \eta_k(\hat{\theta}_k).$$

This simple extension is much more computation intensive than the original procedure, since the $X_{(i)}$'s for $i = k+1, \dots, k+K_n$ must be re-computed for each k , instead of once and for all. It is illustrated in Figure 4, using the same simulated dataset as in the previous Sections. Here the null distribution is defined as the Gaussian $\mathcal{N}(0; \sigma^2)$, and the unknown variance σ^2 is estimated by the usual mean squares: $\hat{\sigma}_k^2 = \frac{1}{n-k} \sum_{i=k+1}^n Y_{(i)}^2$. The estimated number of significant coefficient, $\hat{k}_n = 95$ is still close to its true value, and so is the corresponding standard error estimate $\hat{\sigma}_{\hat{k}_n}^2 = 1.05$.

3.2 Varying window extension

As we have seen previously, the RT procedure depends on a parameter K_n which can be interpreted as a window width, since η_k is a function of $X_{(k+1)}, \dots, X_{(k+K_n)}$. K_n must be smaller than the number of null coefficients, but at the same time not too small, or η_k would become unstable. Hence choosing an appropriate value for K_n may be a hindrance in practice, especially since we want the RT method to be adaptive and depend as little as possible on any form of tuning.

This issue can be avoided by re-defining η_k as a function of $X_{(k+1)}, \dots, X_{(n)}$, thus replacing the fixed width K_n by a varying width $n - k$, which requires no prior tuning. We define the following procedure, starting with the same steps 1. and 2. as in Section 2.1, and adding the following steps:

3. Let κ_n be a lower bound on the number of null coefficients. For $1 \leq k \leq n - \kappa_n$ and $1 \leq j \leq n - k$, compute:

$$\begin{aligned} T_{k,j} &= \sum_{i=k+1}^{k+j} X_{(i)} \\ Q_{k,j} &= \mathbb{E}_k(T_{k,j} | T_{k,n-k}) \\ \eta_k &= \max_{1 \leq j \leq n-k} |T_{k,j} - Q_{k,j}| / \sqrt{n-k}. \end{aligned} \tag{3}$$

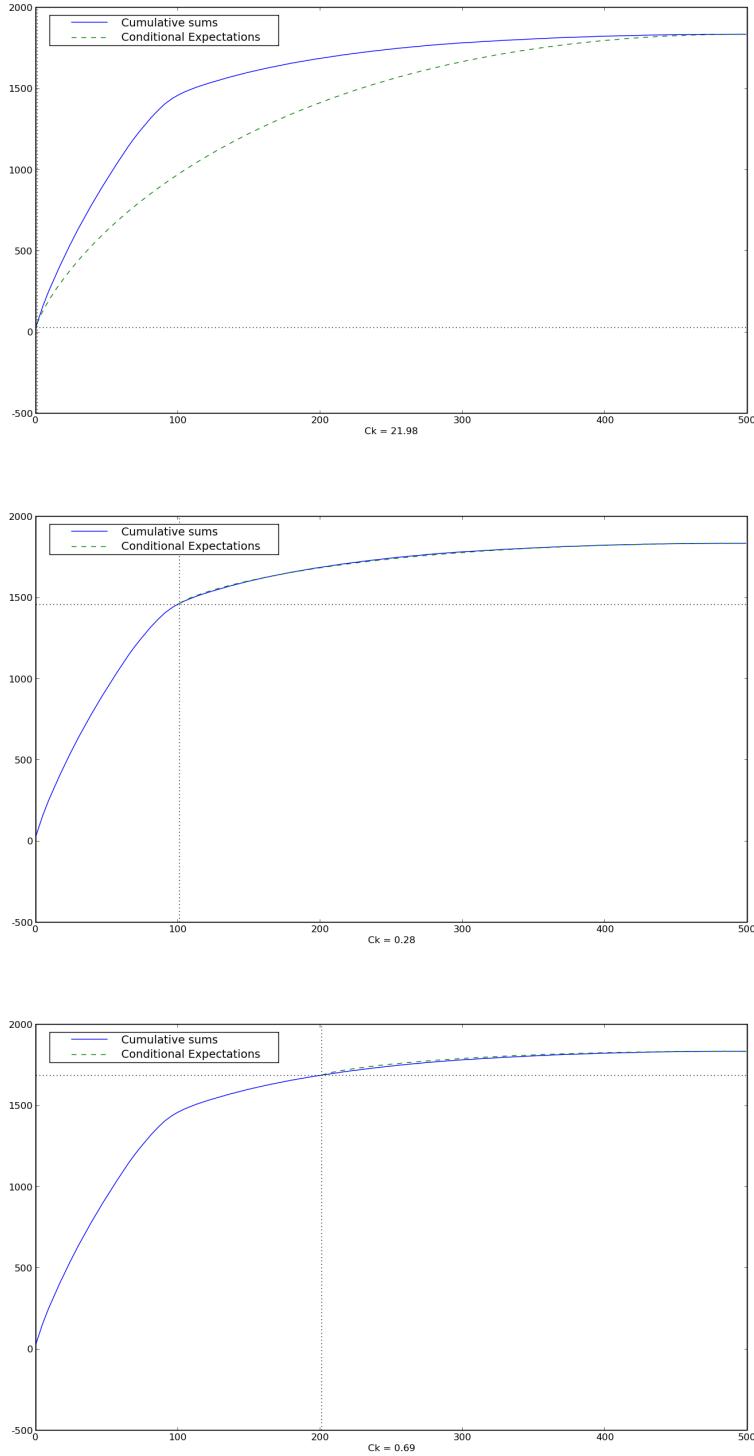


Figure 5: Random threshold with a varying window width: partial cumulative sums $(T_{k,j})_j$ and their conditional null expectations $(Q_{k,j})_j$ for $k = 0$ (top), $k = 100$ (middle), and $k = 200$ (bottom).

4. Let $\hat{k}_n = \operatorname{argmin}_{1 \leq k \leq n - \kappa_n} \eta_k$.

In other terms, η_k would be strictly equal to the test statistic D_n defined in Section 2.1, if the sequence $(X_{(i)})_{1 \leq i \leq n}$ where replaced by the subsequence $(X_{(i)})_{k+1 \leq i \leq n}$, *i.e.*, the set of null terms under $\mathcal{H}_{1,k}$.

Notice that \hat{k}_n is independent from κ_n , as long as η reaches its global minimum on $\{1, \dots, n - \kappa_n\}$. The varying window extension presented here can of course be combined with the unknown distribution extension in Section 3.1.

Figure 5 illustrates the varying window extension on the same simulated dataset as previously. Intuitively, the ‘sliding’ window is replaced by a ‘shrinking’ window, which at first encloses all observations, than progressively reduces in width as the largest observations are left out. Otherwise the same observations as in the fixed-width case hold: when $k = 1$, the cumulative sums $(T_{k,j})_j$ are larger than their conditional expectations $(Q_{k,j})_j$, resulting in a large gap ($\eta_0 = 21.98$). This gap is considerably reduced when $k = 100$ as the $T'_{k,j}$ s and the $Q'_{k,j}$ s become of the same order ($\eta_{100} = 0.28$), then the gap increases again for $k = 200$ as the cumulative sums become smaller than their expected values ($\eta_{200} = 0.69$). Though not shown here, the sequence η_k attains its minimum in $\hat{k}_n = 99$, as in the fixed-window case.

3.3 Asymptotic properties

The estimator of the number of significant coefficients presented in Section 2.2 is consistent. This is the main result in [9], and it can be extended to the varying window setting. We start by recalling the following asymptotic framework:

AF1 There exists $t^* \in (0, 1)$ and a subset $I_{k_n^*}$ of $\{1, \dots, n\}$, with $k_n^* = [t^*n]$ and $|I_{k_n^*}| = k_n^*$, such that $\mu_i \neq 0$ if $i \in I_{k_n^*}$. For all other index, $\mu_i = 0$.

AF2 For any $i \in I_{k_n^*}$, $|\mu_i| \geq \alpha_n$, for a certain sequence (α_n) with $\alpha_n \rightarrow \infty$ (see [9] for further details).

AF3 $\kappa_n/n \rightarrow c$ such that $0 < c < 1 - t^*$.

We then have the following result:

Theorem 2 (Consistency of the random threshold) *Let \hat{k}_n stand for the estimator defined in Section 2.2. Under assumptions AF1, AF2, AF3, and appropriate Von-Mises type conditions on the cdf F_ϵ of the errors (see [9] for further details), \hat{k}_n is consistent in the sense that:*

$$P\left(\left|\frac{\hat{k}_n}{n} - t^*\right| > u_n\right) \rightarrow 0, \quad (4)$$

for any positive decreasing sequence (u_n) such that $\sqrt{n}u_n \rightarrow \infty$.

This result can be refined by deriving an upper bound, which we do not detail here, on the convergence rate of the probability in Equation (4), for a particular choice of sequence (u_n) . Consistency also holds in the unknown distribution case, under a different set of assumptions which we do not recall here, and under the varying window extension, as shown in Appendix A.

This theorem is interesting in that it gives a convergence rate for \hat{k}_n , provided that a minimal signal-to-noise (SNR) ratio is attained, represented by a lower bound α_n on the absolute values $|\mu_i|$ of the non-zero means (assumption **AF2**). Note that, in order for the random threshold (or any other threshold for that matter) to asymptotically separate perfectly null from non-null data, this SNR must necessarily become arbitrarily large as the sample size increases.

However, this theorem provides no clue to what happens when the SNR remains bounded, as we expect to be the case in real-life applications. In the remainder of this paper, our goal is to explore the behaviour of the RT approach in such cases.

4 Simulation study

In order to assess empirically the classification properties of RT, we designed several numerical experiments. Our goal was to compare the binary classification risk of the RT procedure (with both fixed and varying window width), to those obtained by model-based clustering and FDR control techniques. Specifically, we used a mixture-model, estimated via an expectation-maximisation (EM) algorithm [4] to approximate the risk minimizing detection threshold, and the Benjamini-Hochberg (BH) procedure [2] to derive a threshold controlling the FDR at a certain level. We dismissed FWER control techniques as they essentially yield constant thresholds at a given level, and are therefore of little interest when compared to adaptive approaches.

We considered two cases, depending on whether the null distribution F_ϵ was considered as known or not. Note that the BH procedure is based on the p-values $p_i = 1 - F_{|\epsilon|}(|Y_i|)$, hence it requires that F_ϵ be known, whereas this same distribution can be estimated using the EM algorithm. So in order to compare methods on a fair basis, we compared RT to the BH procedure when the null distribution was known, and to mixture model fit otherwise.

4.1 Results with known null distribution

We chose to simulate the X_i directly in the case of a known null distribution. Datasets of $n = 10\,000$ observations were generated, containing each $n_1 = 1\,000$ significant terms. These were sampled from the Gamma distribution $\mathcal{G}(\alpha, \beta)$, where β is a scale parameter, and the remaining 9 000 null terms were sampled from the $\mathcal{E}(1)$ distribution.

Table 1 shows the average classification risks obtained by the different methods over 100 simulated datasets and for different choices of the Gamma distribution parameters. More precisely, we chose to compute the ratio of each attained risk to the lowest achievable (oracle) risk, which makes more sense since perfect classification is in general unattainable.

The binary risk and oracle threshold can be computed as follows. Consider a given dataset $(X_i, Z_i)_i$, where Z_i is a binary variable, equal to 0 if X_i is a null term (sampled from the $\mathcal{E}(1)$ distribution), and 1 if X_i is a non-null term (sampled from the Gamma distribution). Then the overall classification error associated to a given detection threshold t is given by:

$$c(t) = \sum_{Z_i=0} \mathbf{1}_{\{X_i > t\}} + \sum_{Z_i=1} \mathbf{1}_{\{X_i \leq t\}}, \quad (5)$$

	β			β			β			
	1.0	2.0	3.0	1.0	2.0	3.0	1.0	2.0	3.0	
α	5.0	1.88	1.85	1.36	1.6	1.08	1.12	1.31	1.06	1.65
	6.0	2.4	1.66	1.14	1.59	1.04	1.61	1.19	1.33	3.00
	7.0	2.7	1.37	1.09	1.42	1.21	2.91	1.07	2.01	6.02
	FDR 0.01			FDR 0.05			FDR 0.1			

	β			β			
	1.0	2.0	3.0	1.0	2.0	3.0	
α	5.0	1.31	1.15	1.11	1.24	1.13	1.10
	6.0	1.30	1.14	1.14	1.25	1.12	1.14
	7.0	1.27	1.13	1.16	1.23	1.12	1.17
	fix. RT			var. RT			

Table 1: Ratio of binary classification risks with respect to the lowest attainable (oracle) risk for FDR control at different levels (top) and for the RT procedure with fixed ($K_n = 5000$) and variable ($\kappa_n = 5000$) window width (bottom), averaged over 100 simulated datasets.

that is, the sum of type I (false detections) and type II (non detections) errors. The oracle threshold is then chosen to minimize this classification error:

It can be seen that both RT approaches perform in general better than FDR control through the BH procedure, with a slight advantage to the varying window extension. Most importantly, the classification risks they attain is never more than 1.31 and 1.25 times the oracle risk for the fixed and varying window version, respectively. In contrast to these near optimal performances, whatever the chosen level of FDR control, the BH procedure always performs poorly for at least one model, with an average classification risk that rises as high as 6.02 times the optimal one in the worst case.

These results suggest that the RT approach, due to its adaptive nature, is indeed more stable than error rate control techniques, that depend on the choice of a false detection level, as we had anticipated. Moreover, the excellent performance of the RT methods, which attained near optimal risks on the studied datasets, is very encouraging for this approach.

4.2 Results with unknown null distribution

To illustrate the unknown distribution case, we simulated $n = 1000$ observations Y_i , among which $n_1 = 100$ were sampled from the $\mathcal{N}(\mu, \sigma^2)$ distribution with $\mu > 0$ and represented the significant terms, and $n - n_1 = 900$ were sampled from the $\mathcal{N}(0, 1)$ distribution and represented the null terms. We used less observations than in the known distribution case because the methods used in the present case are much more computer-intensive.

We implemented an EM algorithm to estimate a two-class Gaussian mixture model (GMM) from the data, with one zero-mean class to model the null data. As is often the case with iterative algorithms, providing initial values for the model parameters was the main problem we encountered. We found an efficient strategy for doing so,

	σ			σ			σ			
	1.0	2.0	3.0	1.0	2.0	3.0	1.0	2.0	3.0	
μ	1.0	1.03	1.03	1.08	1.03	1.06	1.02	1.03	1.06	1.03
	2.0	1.06	1.03	1.04	1.32	1.13	1.05	1.30	1.12	1.05
	3.0	1.11	1.06	1.04	1.60	1.19	1.08	1.55	1.18	1.08
	GMM fit			fix. RT			var. RT			
	4.01			2.03			1.89			
	GMM fit			fix. RT			var. RT			

Table 2: Top: Ratio of binary classification risks over oracle risk for model-based clustering (left) and for the RT procedure with fixed ($K_n = 500$) and variable ($\kappa_n = 500$) window width (middle and right), averaged over 100 simulated datasets. Bottom: Results for bimodal non-null data (averaged over 100 simulated datasets).

taking advantage of the fact that the negative data contained mostly null terms, and could provide a good initial guess for the null distribution variance and mixture weight. Details of the algorithm are given in Appendix B.

Table 2, top shows the average ratios of the classification risks obtained by the different methods with respect to the oracle risk, over 100 simulated datasets and for different choices of the Gaussian distribution parameters for the significant terms.

All methods performed satisfactorily, yielding close to optimal risks, with model-based clustering performing slightly better than the RT methods. This comes as a little surprise, since in this case the former approach had several advantages: it was based on a parametric model for the significant terms that was precisely the one used to simulate the data, and it explicitly minimized the binary classification risk. In contrast, the RT approach requires no parametric assumption on the distribution of significant terms, and does not explicitly minimize a classification risk, but nevertheless gave good results.

Furthermore, performances of the model-based method can deteriorate when it is based on the wrong assumptions. To illustrate this, we simulated $n = 5000$ observations Y_i among which $n_1 = 950$ were sampled from the $\mathcal{N}(3, 1)$ distribution, $n_2 = 50$ from the $\mathcal{N}(20, 1)$ distribution, and the remaining 4000 from the $\mathcal{N}(0, 1)$ distribution and represented the null data. Consequently, the significant terms had a bimodal distribution. Most of these terms were next to the null mode, and a small number were next to a more distant mode.

This way, we hoped to trick the mixture model, which assumed a unimodal distribution for the significant terms, in detecting only the distant mode, while merging the other one with the null distribution. This is exactly what happened, as can be seen in Table 2, bottom: the mixture-model fit performs significantly worse than the RT approach in this case, the later maintaining a reasonable, though also degraded, classification risk.

Of course it can be argued that such a dataset does not represent a realistic situation; our point here is simply to illustrate the increased robustness of RT due to the fact that it requires no assumptions other than a noise model. It can also be discussed

that an alternative to the simplistic two-class GMM used here would be to allow a variable number of classes, combined with a model selection framework [8, 11, 5]. However, implementing such complex strategies would be non-trivial, especially concerning the algorithm’s initialization. This last issue could be addressed for instance by using stochastic extensions of the EM, such as the stochastic averaging EM (SAEM) [3], in order to reduce dependence on initial values. In contrast to such sophisticated strategies, the simplicity of the RT approach, which requires minimal implementation and virtually no tuning, appears as a key advantage in practice, especially in view of the good performances suggested by this study.

5 Application to fMRI data analysis

We now apply the random threshold approach to functional magnetic resonance imaging (fMRI) data analysis. fMRI is a modality of *in vivo* brain imaging that allows to measure the variations of cerebral blood oxygen levels induced by the neural activity of a subject lying inside a MRI scanner and submitted to a series of stimuli. A sequence of three-dimensional (3D) images of the brain is thus acquired, measuring over time a vascular effect of neural activity known as the blood oxygenation level dependent (BOLD) effect. From the time series recorded in each voxel, and the occurrence times for each stimulus, one may compute an estimate of the BOLD effect of the subject in response to any given stimulus, and more generally to any difference or combination of stimuli (contrast) [6, 16].

Thus, the fMRI data for one subject generally consists in a spatial map of z -scores (Y_1, \dots, Y_p) , where p is the number of voxel in the search volume (which can be as high as 100 000), and Y_i the estimated BOLD effect at voxel i . This map of measures of cerebral activity, also termed *activation map*, is plagued by several sources of uncertainty: the natural variability of brain activity, and the estimation noise induced by the MRI scanner. Thus, model (1) provides a good representation of the activation map (Y_1, \dots, Y_p) , with significant terms corresponding to brain regions involved in the task under study.

In a typical fMRI study however, not one but several subjects are recruited from a population of interest, and scanned while submitted to the same series of stimuli. Activation maps associated with a given contrast are obtained for each subject, as described above, and used as input data for inference at the between-subject level, where the goal is to evidence a general brain activity pattern. Mass univariate, or voxel-based, detection [6] is to date the most widely used approach to address this question. It starts with normalizing individual images onto a common brain template using nonrigid image registration. Next, a t -statistic is computed in each voxel to locally assess mean group effects.

In both single-subject and multi-subject fMRI data analysis, the problem of activation detection can be formulated statistically as that of detecting non-zero means among a collection of observations. The most common approach consists in thresholding a statistical map of brain activity [6]. Multiple-testing techniques are widely used [12, 13], as well as mixture-models. The Gamma-Gaussian mixture model (GGM) is most often used in this context [1]. It uses a Gaussian distribution for null, or inactivated, data, a Gamma distribution for activated data, and a negative Gamma distribution for deactivated data.

As these methods suffer from certain limitations, as discussed in the previous Sections, the RT appears as an appealing alternative in this context. Hence we decided to compare the regions detected by the different approaches, to see if RT succeeded in recovering regions known to be involved in certain well-studied cognitive tasks.

5.1 Data acquisition and preprocessings

We used a real fMRI dataset from the Localizer database [14], involving a cohort of 38 right-handed subjects, and acquired as follows. The participants were presented with a series of stimuli or were engaged in tasks such as passive viewing of horizontal or vertical checkerboards, left or right click after audio or video instruction, computation (subtraction) after video or audio instruction, sentence listening and reading. Events occurred randomly in time (mean inter stimulus interval: 3s), with ten occurrences per event type, and ten event types in total.

Functional images were acquired on a General Electric Signa 1.5T scanner using an Echo Planar Imaging sequence . Each volume consisted of $34 \times 64 \times 64$ mm-thick axial contiguous slices. A session comprised 130 scans. Anatomical T1 weighted images were acquired on the same scanner, with a spatial resolution of $1 \times 1 \times 1.2$ mm³. Finally, the cognitive performance of the subjects was checked using a battery of syntactic and computation tasks.

Single-subject analyses were conducted using SPM5 (<http://www.fil.ion.ucl.ac.uk>). Data were submitted successively to motion correction, slice timing and normalization to the MNI template. For each subject, BOLD contrast images were obtained from a fixed-effect analysis on all sessions. Group analyses were restricted to the intersection of all subjects' whole-brain masks, comprising 43 367 voxels.

We considered the *t*-score maps computed for different contrasts of experimental conditions. These were first converted to *z*-score maps, to obtain approximatively Gaussian statistics in inactivated voxels. Using these maps as input data, we then compared the detection thresholds obtained by Gamma-Gaussian mixture modeling (GGM), fixed-window random thresholding and the varying-window extension, also using the unknown variance extension in both cases (see Section 3.1). For simplicity, we only present here the results obtained for a fixed window equal to $K_n = 15\,000$.

5.2 Individual subject activation map

Our first illustration concerns the activation map of a single subject, for the “sentence–checkerboard” contrast. This contrast subtracts the effect of viewing horizontal and vertical checkerboards from that of reading video instructions, thus allowing to detect brain regions specifically implicated in the reading task.

Figure 6, left, shows an axial slice from the *z*-score map before thresholding. Activations are clearly seen in Wernicke's and Broca's areas (right and upper right), which are known to be involved in language processing (see [15], for instance). The detection threshold found by GGM fit for the *z*-score map (2.03) is much lower than those found by the random threshold procedure, both with a varying window (3.19) and a fixed window (3.33).

The random thresholds with fixed and variable windows yield very similar activation maps in this case, which seem to capture the activated regions seen in the raw map. In contrast, the much lower threshold found by mixture modeling detects several smaller clusters, some of which may be false positives.

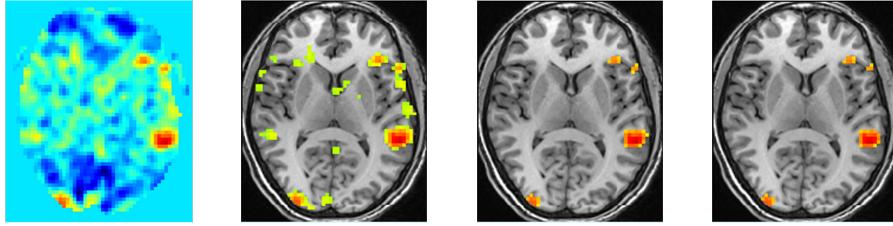


Figure 6: Axial slice from a z -score map for the "sentence–checkerboard" contrast, using a temperature palette for the z -score values. From left to right: Unthresholded, thresholded by GGM model fit, varying-window and fixed-window ($K_n = 15\,000$) random thresholding. Detected activations are shown against the subject's anatomical image.

5.3 Group activation map

In this second example, we consider a group activation map, specifically a map of t -statistics computed from the individual contrast maps of 15 subjects, thus enabling to infer regions of positive mean effects in the parent population. Our choice of limiting the number of subjects, rather than using the whole cohort, was driven by the fact that many fMRI studies are conducted on groups of less than 20 subjects.

We report results for the "calculation–sentences" contrast, which subtracts activations due to reading or hearing instructions from the overall activations detected during the mental calculation tasks. This contrast may thus reveal regions that are specifically involved in the processing of numbers.

Figure 6, left, shows an axial slice from the activation map before thresholding, with clear activations in the bilateral anterior cingulate (upper middle), bilateral parietal (lower left and right) and right precentral (upper right) regions, all known to be involved in number processing [14].

Though sorted in the same order as previously, the varying window random threshold (2.49) is now roughly at equal distances from the threshold found by GGM modeling (1.79) and the fixed window random threshold (3.06).

The three methods detected activations in the regions described above, though the fixed window random threshold seemed to miss some activations, and the GGM approach further detected smaller clusters, some of which may be false positives.

Of course one cannot conclude from these examples only that RT is 'better' at detecting activations than GGM fit. However, the varying window extension successfully detected regions known to be involved in the two cognitive tasks considered here, while avoiding isolated peaks in other regions, which may be part of the background noise. These results suggest that the RT succeeded in capturing only the active regions, while the GGM approach seemed to detect spurious activations.

6 Conclusion

In this paper, we have introduced a simple modification to the random threshold (RT) procedure proposed in [9], to obtain an entirely unsupervised procedure for recovering non null mean terms from a collection of independent random observations, based

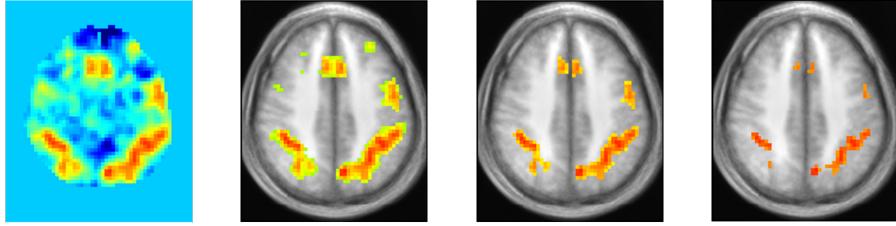


Figure 7: Axial slice from the group activation z -score map for the "calculation–sentence" contrast, using a temperature palette for the z -score values. From left to right: Unthresholded, thresholded by GGM model fit, varying-window and fixed-window random thresholding. Detected activations are shown against the mean anatomical image of all subjects.

solely on a parametric model of the null terms. Our modification, which requires no prior tuning, conserves the consistency properties of the original procedure.

We have implemented all the different versions of the random threshold method in a Python package. This was integrated to the Neuroimaging in Python (NIPY) open-source library, freely downloadable from <http://nipy.sourceforge.net>.

We validated this approach through extensive numerical experiments, and showed that both the original procedure, based on a fixed window-width, and our extension, which uses a variable window, compare favorably to both multiple testing procedures, and model-based clustering, in terms of the binary classification risk, with a slight advantage to our varying window extension. On the vast majority of simulated datasets, the risks achieved were close to the lowest achievable (oracle) risk, whereas each of the other approaches behaved poorly in at least one case.

Thus RT appears as a very promising method for non-ordered model selection whenever no parametric assumptions are available concerning the data distribution. Such methods are needed in many application domains, as we have illustrated in the case of activation detection for fMRI data analysis.

The good classification performances of RT evidenced empirically in our simulations suggest that a promising direction for future research would be to study its properties in the mixture-model setting, and especially its large-sample behaviour. An interesting question to answer would be whether the random threshold converges to a certain limit when the SNR remains constant, and if so, how does this limit compares to the oracle threshold.

References

- [1] C. Beckmann, M. Woolrich, and S. Smith. Gaussian / gamma mixture modelling of ica/glm spatial maps. In *Ninth Int. Conf. on Functional Mapping of the Human Brain.*, 2003.
- [2] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, 57(1):289–300*, 1995.

- [3] B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the em algortihm. *The Annals of Statistics*, 27(1):94–128, 1999.
- [4] A.P. Dempster, A.P. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.
- [5] Bradley Efron, Trevor Hastie, Lain Johnstone, and Robert Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.
- [6] K. J. Friston. *Human Brain Function*, chapter 2, pages 25–42. Academic Press, 1997.
- [7] Y. Ge, S. Dudoit, and T.P. Speed. Resampling-based multiple testing for microarray data analysis. Technical report, Department of Statistics, University of California, Berkeley 2. Division of Biostatistics, University of California, Berkeley 3., 2003.
- [8] T. Hastie, R. Tibshirani, and J. Friedman, editors. *The Elements of Statistical Learning*. Springer Series in Statistics, 2001.
- [9] M. Lavielle and C. Ludeña. Random threshold for linear model selection. *ESAIM: Probability and Statistics*, 2007.
- [10] David J. C. MacKay. *Information theory, inference, and learning algorithms*. Cambridge University Press, 1st edition, June 2003.
- [11] S. Massart. *Concentration Inequalities and Model Selection*. Lecture Notes in Mathematics. Springer, 2003.
- [12] T.E. Nichols and S. Hayasaka. Controlling the Familywise Error Rate in Functional Neuroimaging: A Comparative Review. *Statistical Methods in Medical Research*, 12(5):419–446, 2003.
- [13] M. Perone Pacifico, C. Genovese, I. Verdinelli, and L. Wasserman. False discovery control for random fields. *Journal of the American Statistical Association*, 99(468):1002–1014, 2004.
- [14] P. Pinel, B. Thirion, S. MÃƒriaux, A. Jobert, J. Serres, D. Le Bihan, J.-B. Poline, and S. Dehaene. Fast reproducible identification and large-scale databasing of individual functional cognitive networks. *BMC Neurosci*, 8(1):91, Oct 2007.
- [15] Cathy J. Price. The anatomy of language: contributions from functional neuroimaging. *Journal of Anatomy*, 197(3):335–359, October 2000.
- [16] K.J. Worsley, C.H. Liao, J. Aston, V. Petre, G.H. Duncan, F. Morales, and A.C. Evans. A general statistical analysis for fMRI data. 15(1):1–15, January 2002.

A Proof of the Theorem 2 for the varying window extension

Following [9], we first recall some notations. Set $U_i = Y_i$ for $i \in I_{k_n^*}$ and $v_i = Y_i$ for $i \notin I_{k_n^*}$; notice that (v_i) is a sample from the distribution F_e . Let $(u_{(i)})_{1 \leq i \leq k_n^*}$ and $(v_{(i)})_{1 \leq i \leq n-k_n^*}$ be the sequences $(|U_i|)$ and $(|v_i|)$ in decreasing order. Let Ω_n be the subset of Ω where $v_{(1)} < \alpha_n/2$ and $u_{(k_n^*)} > \alpha_n/2$.

A first lemma in [9] shows that $P(\Omega_n) \rightarrow 1$, *i.e.*, the collections $(u_{(i)})$ and $(v_{(i)})$ are stochastically in order with high probability. The proof can then be restricted to Ω_n .

Now, let $\mathbb{E}_k(T_{k,j})$ and $Q_{k,j}$ be defined as in Equation (3). Using Proposition 1, we have:

$$\begin{aligned}\mathbb{E}_k(T_{k,j}) &= j\left(1 + \sum_{i=j+1}^{n-k} 1/i\right); \\ Q_{k,j} &= \frac{\mathbb{E}_k(T_{k,j})}{\mathbb{E}_k(T_{k,n-k})} T_{k,n-k} \\ &= B_{k,j,n} T_{k,n-k}.\end{aligned}$$

Also, let $a_i = \mathbb{E}_0(Z_{(i)}) = \sum_{\ell=i}^n 1/\ell$. Equation (4) can be shown separately for $k > k_n^*$ and $k < k_n^*$. Since the two cases are treated similarly, we will restrict ourselves here to the case $k > k_n^*$. On Ω_n :

$$\begin{aligned}T_{k,j} - Q_{k,j} &= T_{k,j} - B_{k,j,n} T_{k,n-k} \\ &= (T_{k,j} - \mathbb{E}_{k_n^*}(T_{k,j})) - B_{k,j,n} (T_{k,n-k} - \mathbb{E}_{k_n^*}(T_{k,n-k})) \\ &\quad + \mathbb{E}_{k_n^*}(T_{k,j}) - B_{k,j,n} \mathbb{E}_{k_n^*}(T_{k,n-k}) \\ &= R_{k,j} + S_{k,j}.\end{aligned}$$

Thus $T_{k,j} - Q_{k,j}$ is decomposed into a random part $R_{k,j}$ and a deterministic part $S_{k,j}$. Over Ω_n , $R_{k,j}$ is a function of $v_{(k)}, \dots, v_{(n-k_n^*)}$. Before going further, we now recall the following result:

Let $Z_{(1)} \geq \dots \geq Z_{(n)}$ be an ordered sequence of independent $Exp(1)$ random variables. For $1 \leq j \leq n$, let $T_j = \sum_{i=1}^j Z_{(i)}$. Introduce for $t \in [0, 1]$ the random process $d_n(t) = T_{[nt]} - \mathbb{E}(T_{[nt]}|T_n)$. Then it is shown in [9] that $\frac{1}{\sqrt{n}}d_n(t)$, as a process indexed on $t \in [0, 1]$, converges in distribution to a certain zero mean Gaussian process Δ .

To use this result, let $k = [tn]$ and $j = [sn]$, for $0 < t < 1 - c$ and $0 < s < 1 - t^* - t$, for c in **[AF3]**. Then $\frac{1}{\sqrt{n-k}}(T_{k,j} - Q_{k,j})\mathbf{1}_{\Omega_n} = \frac{1}{\sqrt{n-k}}(T_{[tn],[sn]} - Q_{[tn],[sn]})\mathbf{1}_{\Omega_n}$, as a process indexed by $(t, s) \in (0, 1)^2$, converges in distribution to the zero-mean Gaussian process:

$$\Gamma_{t,s} = \sqrt{\frac{1-t^*}{1-t}} \left[\Delta \left(\frac{t+s-t^*}{1-t^*} \right) - \Delta \left(\frac{t-t^*}{1-t^*} \right) \right].$$

similarly, $\frac{1}{\sqrt{n-k}}B_{k,j,n}\mathbb{E}_{k_n^*}(T_{k,n-k})\mathbf{1}_{\Omega_n}$ converges in distribution to another zero-mean Gaussian process, and so does their sum, $\frac{1}{\sqrt{n-k}}R_{k,j}\mathbf{1}_{\Omega_n}$.

On the other hand,

$$S_{k,j} = \sum_{i=1}^{k-k_n^*} (a_{i+j} - a_i + B_{k,j,n}(a_{i+n-k} - a_i)), \quad (6)$$

so that there exists a constant $\gamma > 0$, which depends on c in [AF3], such that for all $n \geq 1$, $k_n^* < k \leq n - K_n$, we have $\sup_{1 \leq j \leq n-k} |S_{k,j}| \geq \gamma(k - k_n^*)$. Finally we use the following inequality:

$$\mathbb{P}_{k_n^*}(\hat{k}_n - k_n^* > nu_n) \leq \mathbb{P}(\eta_{k_n^*} > \inf_{k-k_n^* > nu_n} \eta_k).$$

From Equation (6), $S_{k_n^*,j} = 0$, hence it follows that:

$$\begin{aligned} \sqrt{n - k_n^*} \eta_{k_n^*} &= \sup_{1 \leq j \leq n - k_n^*} |R_{k_n^*,j} + S_{k_n^*,j}| \\ &= \sup_{1 \leq j \leq n - k_n^*} |R_{k_n^*,j}| \\ &\leq \sup_{k \geq k_n^*} \sup_{1 \leq j \leq n - k} |R_{k,j}|. \end{aligned}$$

On the other hand,

$$\begin{aligned} \sqrt{n - k} \inf_{k-k_n^* > nu_n} \eta_k &= \inf_{k-k_n^* > nu_n} \sup_{1 \leq j \leq n - k} |R_{k,j} + S_{k,j}| \\ &\geq \inf_{k-k_n^* > nu_n} \sup_{1 \leq j \leq n - k} |S_{k,j}| - \sup_{k \geq k_n^*} \sup_{1 \leq j \leq n - k} |R_{k,j}|, \end{aligned}$$

so that we have:

$$\begin{aligned} \mathbb{P}_{k_n^*}(\hat{k}_n - k_n^* > nu_n) &\leq \mathbb{P}(C \sup_{k \geq k_n^*} \sup_{1 \leq j \leq n - k} |R_{k,j}| \geq \inf_{k-k_n^* > nu_n} \sup_{1 \leq j \leq n - k} |S_{k,j}|) + \mathbb{P}(\Omega_n^c) \\ &\leq \mathbb{P}(C \sup_{k \geq k_n^*} \sup_{1 \leq j \leq n - k} |R_{k,j}| \geq \gamma nu_n) + \mathbb{P}(\Omega_n^c), \end{aligned}$$

where C is a constant which depends on c in [AF3]. This last probability vanishes as n goes to infinity, due to the weak convergence of $R_{k,j} \mathbf{1}_{\Omega_n}$. \square

B Details of the EM algorithm for the two-class GMM with a zero-mean class

We consider the following model:

$$\begin{aligned} Y_i | Z_i = j &\stackrel{iid}{\sim} \mathcal{N}(\mu_j, \sigma_j^2), \quad i = 1, \dots, n, \quad j = 0, 1 \\ Z_i &\stackrel{iid}{\sim} \mathcal{B}(1, p_1), \end{aligned} \quad (7)$$

where $\mu_0 = 0$, and p_j represents the proportion of data in class j , so that the vector of model parameters is: $\theta = (p_0, \mu_1, \sigma_0, \sigma_1)$.

Having initialized θ to $\theta^{(0)}$, the EM algorithm alternates the following steps:

E-step. Compute the conditional law of the indicator variable Z_i at step t , that is, the Bernoulli defined by:

$$\begin{aligned}\mathbb{P}(Z_i = j|Y_i, \theta^{(t)}) &= \frac{f(Y_i|Z_i = j, \theta^{(t)})p_j^{(t)}}{\sum_j f(Y_i|Z_i = j, \theta^{(t)})p_j^{(t)}} \\ &:= p_{ij}^{(t)}\end{aligned}\quad (8)$$

M-step. Update the estimate of model parameters by maximizing the conditional expectation of the complete log-likelihood, yielding:

$$\theta^{(t+1)} = \arg \max_{\theta} \mathbb{E} \left[\sum_i \log f(Y_i|Z_i, \theta) | \mathbf{Y}, \theta^{(t)} \right],$$

the expectation being taken with respect to the conditional distribution of the indicator variables Z_i computed in the previous step. This yields:

$$\begin{aligned}p_j^{(t+1)} &= \frac{\sum_i p_{ij}^{(t)}}{n}; \\ \mu_1^{(t+1)} &= \frac{\sum_i p_{ij}^{(t)} Y_i}{\sum_i p_{ij}^{(t)}}; \\ \sigma_j^{2(t+1)} &= \frac{\sum_i p_{ij}^{(t)} (Y_i - \mu_j^{(t)})}{\sum_i p_{ij}^{(t)}}.\end{aligned}\quad (9)$$

Note that, throughout the iterations, $\mu_0^{(t)} \equiv 0$.

Initialization. An initial guess for σ_0^2 is provided by the negative data, which consists mainly of null data:

$$\sigma_0^{2(0)} = \#\{Y_i < 0\}^{-1} \sum_{Y_i > 0} Y_i^2.$$

Then, we use a kernel estimate of the data density:

$$\hat{f}(x) = \frac{1}{n} \sum_i \frac{K\left(\frac{x-Y_i}{h_n}\right)}{h_n}, \quad (10)$$

for a symmetric, positive kernel K , and a bandwidth h_n . In practice, we used the Gaussian kernel $K(x) = 2\pi^{-1/2}e^{-x^2/2}$, and $h_n = \sqrt{n}$.

By identifying the null mode of the data density kernel estimate to the null component of the mixture model, we then obtained an initial guess for the mixture weights:

$$p_0^{(0)} = \hat{f}(0) \sqrt{2\pi\sigma_0^{2(0)}}.$$

Finally, the conditional law of the indicator variables where approached by:

$$p_{i1}^{(0)} = \min \left\{ 1; \frac{p_0^{(0)} \exp\{-Y_i^2/2\sigma_0^{2(0)}\}}{\sqrt{2\pi\sigma_0^{2(0)}}} \right\}.$$

These initial guesses are used to derive initial model parameter values $\theta^{(1)}$, via the M-step described above, for $t = 0$.